Mathematical Probability (STAT2003) Notes

Limao Chang

Contents

Preface

These notes based on the 2022 offering of *Mathematical Probability (STAT2003)*, and so the material from this offering may differ from past or future offerings. Extension content and content that was not explicitly covered in the lectures/tutorials have been labelled as such.

The source LaTeX files can be found at https://github.com/LimaoC/course-notes.

Chapter 1

Random Experiments and Probability Models

1.1   Random Experiments

A random experiment is an experiment whose outcome is not known in advance. A random experiment can (theoretically) be repeated infinitely many times, and has a well-defined set of outcomes.

To specify a random experiment, we require a *sample space*, a collection of *events*, and a way of assigning *probabilities* to these events.

1.2   Events

Definition 1.2.1 (Sample space)**.** The *sample space* of an random experiment is the set Ω of all possible outcomes in the experiment. Outcomes may sometimes be referred to as *sample points*.

Definition 1.2.2 (Event)**.** An *event* is any set of outcomes, that is, any subset of Ω. Events are usually denoted by capital letters, e.g. A, B, C. An event A is said to *occur* if the outcome of the random experiment is one of the elements in A.

Since events are just subsets of Ω, they can be combined in the same way that sets can be combined to form new events. Let A and B be two arbitrary events. Then:

- The *union* of A and B, denoted $A \cup B$, is the event that A or B occur, and consists of all outcomes in A or in B.

- The *intersection* of A and B, denoted $A \cap B$, is the event that A and B occur, and consists of all outcomes in both A and B.

- More generally, if $\{A_i\}$ is a collection of events, then $\cup_i A_i = A_1 \cup A_2 \cup \cdots$ is the event that A_i occurs for at least one value of i, and $\cap_i A_i = A_1 \cap A_2 \cap \cdots$ is the event that A_i occurs for all values of i.

- If all the outcomes in A are also in B, then A is a *subset* of B, denoted $A \subset B$. We also say that A *implies* B (since if A occurs, then B must also occur).

- A and B are said to be *disjoint* if they have no outcomes in common (i.e. they cannot both occur), that is $A \cap B = \varnothing$. More generally, a collection $\{A_i\}$ of events is said to be *mutually exclusive* if each pair of events $(A_i, A_j), i \neq j$ is disjoint.

If A is an event, then the event that A *does not occur*, denoted A^c, consists of all outcomes in Ω that are not in A.

The empty set $\varnothing$ is called the *impossible event* (as it never occurs), and its complement Ω is called the *certain event* (as it always occurs).

Theorem 1.2.3 (De Morgan's Laws). *Let $\{A_i\}$ be a collection of events. Then*

$$\left(\bigcup_i A_i\right)^c = \bigcap_i A_i^c \quad and \quad \left(\bigcap_i A_i\right)^c = \bigcup_i A_i^c.$$

1.3   Probability

Recall from §1.1 that for a random experiment, we require a way of assigning probabilities to events. We do this by defining a function $\mathbb{P}$ such that, for any event A, $\mathbb{P}(A)$ can be interpreted as "the probability of A occurring". There are several properties for $\mathbb{P}$ that we should have, such as:

- $\mathbb{P}(\Omega) = 1$, since Ω always occurs.

- $\mathbb{P}(\varnothing) = 0$, since $\varnothing$ never occurs.

- $0 \leqslant \mathbb{P}(A) \leqslant 1$ for any event A, since A either never occurs, always occurs, or sometimes occurs.

- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, for any event A, since either A occurs or A doesn't occur.

We define $\mathbb{P}$ via three fundamental probability axioms, which allow us to derive all other properties of probability. Some examples are demonstrated below.

Definition 1.3.1 (Kolmogorov axioms). The function $\mathbb{P}$ is called a *probability measure* if it satisfies:

(a) For any event A, $\mathbb{P}(A) \geqslant 0$.

(b) $\mathbb{P}(\Omega) = 1$.

(c) If $\{A_i\}$ is a collection of mutually exclusive events, then

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots = \sum_i \mathbb{P}(A_i).$$

Example 1.3.2. To obtain $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, use (b) to get $\mathbb{P}(A \cup A^c) = 1$, (c) to get $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$, and rearrange. □

Example 1.3.3. To obtain $\mathbb{P}(\varnothing) = 0$, note $\varnothing^c = \Omega$, so use the previous example to get $\mathbb{P}(\varnothing) = 1 - \mathbb{P}(\Omega) = 1 - 1 = 0$. □

Example 1.3.4. For events A, B such that $A \subset B$, we have $\mathbb{P}(A) \leqslant \mathbb{P}(B)$. To show this, note $B = A \cup (A^c \cap B)$, where A and $A^c \cap B$ are disjoint (this can be visually confirmed on the Venn diagram below). Then use (c) to get $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(A^c \cap B)$, and then, since $\mathbb{P}(A^c \cap B) \geqslant 0$ by (a), we have $\mathbb{P}(B) \geqslant \mathbb{P}(A)$ as required.


□
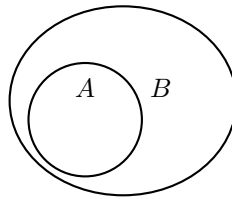
Example 1.3.5. For arbitrary events A, B, we have $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$. Note that any event A may be written as the union of two disjoint events via $A = (A \cap B^c) \cup (A \cap B)$ (this is a specific case of the Law of Total Probability). Hence we can write

$$\mathbb{P}(A) = \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) \quad and \quad \mathbb{P}(B) = \mathbb{P}(B \cap A^c) + \mathbb{P}(B \cap A)$$

by (c). The event $A \cup B$ can be written as the union of three disjoint events, $A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)$, and so we can write

$$\mathbb{P}(A \cup B) = \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

upon substituting the expressions for $\mathbb{P}(A)$ and $\mathbb{P}(B)$.

$\square$
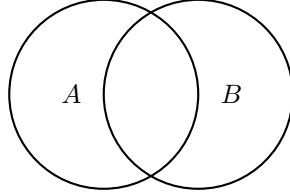
Lemma 1.3.6 (Boole's inequality)**.** *For a countable collection of events $A_1, A_2, \ldots$, we have*

$$\mathbb{P}\left(\bigcup_i A_i\right) \leqslant \sum_i \mathbb{P}(A_i).$$

Proof. Define $B_1 = A_1$ and for $i > 1$, $B_i = A_i \cap A_{i-1}^c \cap \cdots \cap A_1^c$. Then we make use of three observations, namely that

(a) The events $\{B_i\}$ are mutually exclusive by construction, so by the third probability axiom, $\mathbb{P}(\cup_i B_i) = \sum_i \mathbb{P}(B_i)$.

(b) For $n \geqslant 1$, $B_i \subseteq A_i$, since each B_i is always A_i intersected with some other event.

(c) For $i \geqslant 1$, $\cup_i B_i = \cup_i A_i$, since all the complements are "cancelled out" by prior B_i terms.

Then we have

$$\mathbb{P}\left(\bigcup_i A_i\right) = \mathbb{P}\left(\bigcup_i B_i\right) = \sum_i \mathbb{P}(B_i) \leqslant \sum_i \mathbb{P}(A_i),$$

on applying (c), then (a), then (b). $\square$

In the case that Ω is a finite set with all outcomes of Ω being equally likely (that is, they all have the same probability), then, for any event A,

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|},$$

where $|A|$ and Ω are the number of outcomes in A and Ω respectively. This is known as the **equilikely principle**, and can be shown to follow from the third probability axiom:

Proof. Since Ω is a finite set, we can write $\Omega = \{\omega_1, \ldots, \omega_n\}$, where n is the cardinality of Ω, that is $n = |\Omega|$. We can express Ω as a union of disjoint events via $\Omega = \{\omega_1\} \cup \cdots \cup \{\omega_n\}$, and apply the third probability axiom to yield

$$1 = \mathbb{P}(\Omega) = \sum_{i=1}^{n} \mathbb{P}(\{\omega_i\}) \implies \mathbb{P}(\{w_i\}) = \frac{1}{n},$$

since each outcome is equally likely. Thus if A is an event consisting of $k \leqslant n$ outcomes, take wlog the k outcomes to be $\omega_1, \ldots, \omega_k$. Then similarly, write $A = \{\omega_1, \ldots, \omega_k\} = \{\omega_1\} \cup \cdots \cup \{\omega_k\}$, and hence

$$\mathbb{P}(A) = \sum_{i=1}^{k} \mathbb{P}(\{\omega_i\}) = \frac{k}{n}.$$

$\square$

In this case of a finite sample space Ω with all outcomes being equally likely, probabilities can often be calculated by means of counting arguments, which will be discussed in §1.4.

1.4 Counting

Say we have n balls labelled $1, \ldots, n$ placed in an urn and we wish to pick $k \leqslant n$ balls from the urn. Then there are 4 scenarios to consider:

(i) Drawing with replacement, where order matters. In this case, there are n^k ways to choose k balls from the urn, since there are n choices for each of the k balls, so we have $\underbrace{n \times \cdots \times n}_{k \text{ times}} = n^k$.

(ii) Drawing without replacement, where order matters. In this case, there are $^nP_k = \frac{n!}{(n-k)!}$ ways to choose k balls from the urn, since there are n choices for the first ball, $n-1$ choices for the second ball, and so on, with $n-k+1$ choices for the kth ball, so we have $n \times (n-1) \times \cdots \times (n-k+1) =: {}^nP_k$.

(iii) Drawing with replacement, where order doesn't matter. In this case, there are $\binom{n+k-1}{k} = \frac{(n+k-1)!}{k!(n-1)!}$ ways to choose k balls from the urn.

(iv) Drawing without replacement, where order doesn't matter. In this case, there are $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ ways to choose k balls from the urn, since there are $\frac{n!}{(n-k)!}$ ways when considering order (from (ii)), and $k!$ possible orders (k choices for the 1st ball, $k-1$ choices for the second ball, ...), so we divide $\frac{n!}{(n-k)!}$ by $k!$ as we don't care about order.

Remark. (iii) is extension material. The explanation of it below uses the *stars and bars* method in combinatorics.

For an explanation of (iii), we proceed with an example. Consider rolling an n-sided dice k times without caring about order. If we take $n = 3$ and $k = 2$, there are 6 possible outcomes:

$$\{1,1\}, \{1,2\}, \{1,3\}, \{2,2\}, \{2,3\}, \{3,3\}.$$

Imagine we then have $n = 3$ boxes, and on each roll of the die, we put a star $*$ into the box corresponding to that roll; for example, if we roll a 2, we would put a star in the second box. Representing the boundaries of each box as a bar $|$, those same 6 outcomes can be expressed like so:

$$|**|\ |\ |, \qquad |*|*|\ |, \qquad |*|\ |*|, \qquad |\ |**|\ |, \qquad |\ |*|*|, \qquad |\ |\ |**|,$$

where whitespaces have been added for readability. For example, $\{1,1\}$ maps to $|**|\ |\ |$ since we rolled two 1's, zero 2's, and zero 3's, so there are two stars in the first box.

The key observation is that the number of outcomes is equivalent to the number of ways we can arrange $k = 2$ stars and $n - 1 = 2$ bars (there are actually $n + 1 = 4$ bars in the diagrams above, but the two outer bars are always fixed, so we can ignore them). We can think of this as choosing k spots to put the stars, and then putting bars in the remaining $n - k$ spots. There are $n + k - 1$ total spots, so from (iv), this is given by $\binom{n+k-1}{k}$.

1.5 Conditional Probability and Independence

Definition 1.5.1 (Conditional probability)**.** Given two events A and B, we define the *conditional probability that A occurs given that B occurs* by $\mathbb{P}(A|B)$ (in other words, the probability of A given B), where

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Note the above is defined only when $\mathbb{P}(B) > 0$. This intuitively makes sense, since if $\mathbb{P}(B) = 0$, then it does not make sense to talk about any event that depends on B occurring.

Definition 1.5.2 (Independent events)**.** Two events, A and B, are said to be *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

We can extend the idea of independence to more than two events via the following definition.

Definition 1.5.3 (Mutual independence)**.** Let $\{A_i\}$ be a collection of events. They are said to be *pairwise independent* if, for all $i \neq j$, $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$. They are said to be *triplewise independent* if, for all distinct i, j, k, $\mathbb{P}(A_i \cap A_j \cap A_k) = \mathbb{P}(A_i)\mathbb{P}(A_j)\mathbb{P}(A_k)$. We can similarly have *quadruplewise independence* and so on. More generally, they are said to be *mutually independent* if they are pairwise independent and triplewise independent and quadruplewise independent and so on.

Remark. If we have two mutually exclusive events A and B such that $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$, they can *not* be independent, since mutual exclusivity implies $\mathbb{P}(A \cap B) = \mathbb{P}(\varnothing) = 0$, but $\mathbb{P}(A)\mathbb{P}(B) > 0$. This is reasonable, since if we know that A has occurred, then B cannot have occurred (and vice versa). In short, independence $\neq$ mutual exclusivity.

Example 1.5.4 (Series system)**.** Consider a system of two components connected in series. The system is working if and only if both components are working.
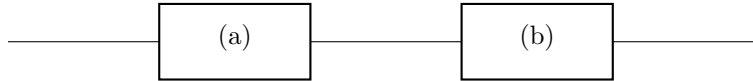
Let A be the event that component (a) is working, and B be the event that component (b) is working, and assume that the two components are independent of each other. Then the probability that the system is working is given by $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ by independence.

Example 1.5.5 (Parallel system)**.** Consider a system of two components connected in parallel. The system is working if and only if at least one component is working.
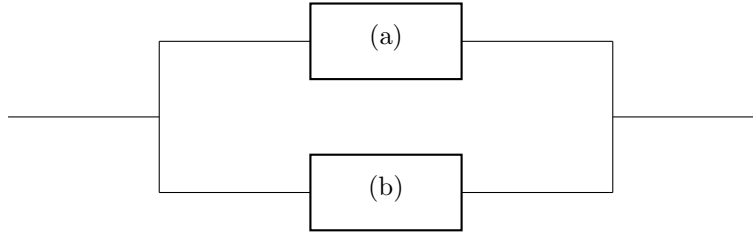
As in the previous example, let A be the event that component (a) is working, and B be the event that component (b) is working, and assume that the two components are independent of each other. Then the probability that the system is working is given by $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B)$.

Reliability systems are covered more in-depth in chapter 7.

Theorem 1.5.6 (The Law of Total Probability)**.** *Let* $B_1, B_2, \ldots$ *form a parition of* Ω *such that* $\mathbb{P}(B_i) > 0$ *for at least one value of* i*. Then, for any event* A,
$$\mathbb{P}(A) = \sum_i \mathbb{P}(A|B_i)\mathbb{P}(B_i),$$
where we can see that the ith term is absent from the sum if $\mathbb{P}(B_i) = 0$.

We will most often use the simplest case of the law, where our partition is $\{B, B^c\}$: Let A and B be arbitrary events such that $0 < \mathbb{P}(B) < 1$. Then
$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c).$$

Proof. We provide a proof for the simple case above (which can be extended to the general theorem statement). Note that we can write A as the union of disjoint events via
$$A = A \cap \Omega = A \cap (B \cup B^c) = (A \cap B) \cup (A \cap B^c).$$

Then using the third probability axiom and the definition of conditional probability,
$$\mathbb{P}(A) = \mathbb{P}((A \cap B) \cup (A \cap B^c)) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c).$$

$\square$

An important corollary of the Law of Total Probability is *Bayes' Theorem*, which is stated below.

Theorem 1.5.7 (Bayes' Theorem). *If A and B are arbitrary events such that $\mathbb{P}(B) > 0$ and $0 < \mathbb{P}(A) < 1$, then*

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}.$$

Alternatively, we may write

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)}.$$

Proof. From the definition of conditional probability, we have

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}.$$

To obtain the alternative expression, use the Law of Total Probability to rewrite $\mathbb{P}(A)$ as

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c).$$

$\square$

Chapter 2

Random Variables and Probability Distributions

Chapter 3

Generating Random Variables on a Computer

Chapter 4

Joint Distributions

Chapter 5

Functions of Random Variables and Limit Theorems

Chapter 6

Markov Chains

Chapter 7

Reliability